

Axiomatic Ranking of Network Role Similarity*

Ruoming Jin Victor E. Lee Hui Hong
 Department of Computer Science
 Kent State University, Kent, OH, 44242, USA
 {jin,vlee,hhong}@cs.kent.edu

ABSTRACT

A key task in social network and other complex network analysis is role analysis: describing and categorizing nodes according to how they interact with other nodes. Two nodes have the same role if they interact with *equivalent* sets of neighbors. The most fundamental role equivalence is automorphic equivalence. Unfortunately, the fastest algorithms known for graph automorphism are nonpolynomial. Moreover, since exact equivalence may be rare, a more meaningful task is to measure the role *similarity* between any two nodes. This task is closely related to the structural or link-based similarity problem that SimRank attempts to solve. However, SimRank and most of its offshoots are not sufficient because they do not fully recognize automorphically or structurally equivalent nodes. In this paper we tackle two problems. First, what are the necessary properties for a role similarity measure or metric? Second, how can we derive a role similarity measure satisfying these properties? For the first problem, we justify several axiomatic properties necessary for a role similarity measure or metric: range, maximal similarity, automorphic equivalence, transitive similarity, and the triangle inequality. For the second problem, we present RoleSim, a new similarity metric with a simple iterative computational method. We rigorously prove that RoleSim satisfies all the axiomatic properties. We also introduce an iceberg RoleSim algorithm which can guarantee to discover all pairs with RoleSim score no less than a user-defined threshold θ without computing the RoleSim for every pair. We demonstrate the superior interpretative power of RoleSim on both synthetic and real datasets.

1. INTRODUCTION

In social science, it is well-established that individual agents tend to play roles or assume positions within their interaction network. For instance, in a university, each individual can be classified into the position of faculty member, administration, staff, or student. Each role may be further partitioned into sub-roles: faculty may be further classified into tenure-track or non-tenure-track positions, etc. Indeed, role discovering is a major research subject in classical social science [45]. Interestingly, recent studies have found not only do roles appear in other types of networks, including food webs [30], world trade [16], and even software systems [9], but also roles can help predict node functionality within their domains. For instance, in a protein interaction network, proteins with similar roles tend to serve similar metabolic functions. Thus, if we know the function of one protein, we can predict that all other proteins having a similar role would also have similar function [18].

Role is complementary to network clustering, a major tool in analyzing network structures. Network clustering attempts to decompose a network into densely connected components. It produces a high level structural model consisting of a small number of “cluster-nodes” and the “super-edges” between these cluster-nodes. Since its goal is to minimize the number of edges (interactions) between clusters, it will result in strong interactions between nodes within each cluster. Given this, the clustering scheme inevitably overlooks and over-simplifies the interaction patterns of each node. For instance, each node in a cluster may take very different “roles”: some of them may serve as the core of the clusters, some may be peripheral nodes, and some serve as the connectors to link between clusters. Indeed, those nodes with similar or same roles may not even directly link to each other as they may simply share similar interaction patterns. Furthermore, even when a network lacks modularity structure, for instance, a hierarchical structure, roles can still be applied for characterizing the interaction patterns of each node. To sum, “roles” provide an orthogonal abstraction for simplifying and highlighting the complex interactions among nodes.

A central question in studying the roles in a network system is how to define *role similarity*. In particular, how can we rank two nodes’ role similarity in terms of their interaction patterns? Despite its vital importance for network analysis and decades of work by social scientists, joined recently by computer scientists, no satisfactory metric for role similarity has yet emerged. A key issue is the encapsulation of graph automorphism (and its generalization) into a role similarity metric: *if two nodes are automorphically equivalent, then they should share the same role and their role similarity should be maximal*. From a network topology viewpoint, automorphic nodes have equivalent surroundings, so one can replace the other. Figure 1 illustrates a graph with nodes $S1$ and $J1$ being automorphically equivalent. Automorphism can be further generalized in terms of *coloration*: assuming each node is assigned a color, then two nodes are equivalent if their neighborhoods consist of the same color spectrum [12].

Traditionally, the social science community has approached role analysis by defining suitable mathematical equivalence relations so that nodes can be partitioned into equivalence classes (roles). An essential property of these equivalences is that they should positively confirm automorphic equivalence, i.e., if any two nodes are automorphic, then they are role-equivalent. (The converse is not necessarily true.) Automorphism confirmation is an instance of verifying a solution, which is often algorithmically less complex than discovering a solution. Therefore, even though there is no known polynomial-time algorithm for discovering graph automorphism¹, role equivalence algorithms [3, 5, 40]

*A revised version of this paper will be published for KDD’11, August 2011.

¹The computational complexity of graph isomorphism and automorphism are still unproven to be either P or NP – *Complete*.

can still guarantee to satisfy the aforementioned automorphism confirmation property. These equivalence rules also directly correspond to the aforementioned coloration.

However, by relying on strict equivalence rules, these role modeling schemes can produce only binary similarity metrics: two nodes are either equivalent (similarity = 1) or not (similarity = 0). In real-world networks, usually only a very small portion of the node-pairs would satisfy an equivalence criteria [31] and among those, many are simply trivially equivalent (such as singletons or children of the same parent). In addition, strict rule-based equivalence is not robust with respect to network noise, such as false-positive or false-negative interactions. Thus, it is desirable in many real world applications to rank node-pairs by their degree of similarity or provide a real-valued node similarity *metric*.

Several recent research works have proposed to measure real-valued structural similarity or to rank nodes' similarity based on their interaction patterns [19, 22]. SimRank [19] is one of the best-known such measures. It generates a node similarity measure based on the following principle: "two nodes are similar if they link to similar nodes". Mathematically, for any two different nodes x and y , SimRank computes their similarity recursively according to the average similarity of all the neighbor pairs (a neighbor of x paired with a neighbor of y). A single node has self-similarity value 1. This is equivalent to the probability that two simultaneous random walkers, starting at x and y , will eventually meet. Most of the existing node structural similarity measures [1, 13, 23, 48, 49, 50] are variants of SimRank. Though SimRank seems to capture the intuition of the above recursive structural similarity, its random walk matching does not satisfy the basic graph automorphism condition. For example, in Figure 1, though $S1$ and $J1$ are automorphically equivalent, SimRank assigns them a value of 0.226. We discuss this further in Section 3.2. To our best knowledge, there is no available real-valued structural similarity measure satisfying the automorphic equivalence requirement. Since automorphic equivalence is a pivotal characteristic of the notion of role, its lack disqualifies these existing measures from serving as authentic role similarity measures. Here is a paradox: SimRank and its variants seem to implement the recursive structural similarity definition of automorphic equivalence (two nodes are similar if they link to similar nodes), yet they do not produce desired results (to assign value 1 to those pairs).

Thus we have an open problem: *Can we derive a real-valued role similarity measure or ranking which complies with the automorphic equivalence requirement?* In this paper, we develop the first real-valued similarity measure to solve this problem. In addition, our measure is also a metric, i.e., it satisfies the triangle inequality. The key feature of our role similarity measure is a weighted generalization of the *Jaccard coefficient* to measure the neighborhood similarity between two nodes. Unlike SimRank, which considers the average similarity among all possible pairings of neighbors, our measure considers only those pairs in the optimal matching of their two neighbor sets which maximizes the targeted similarity function. We show this approach successfully resolves the aforementioned SimRank paradox.

2. ROLE EQUIVALENCE

In social network analysis, the traditional approach for formalizing roles and role groups is to define an equivalence relation and to partition the actors into equivalence classes. Actors who fulfill the same role are equivalent. Over the years, four definitions, offering different degrees of strictness, have stood out. These four, in decreasing strictness order, are structural equivalence, automorphic equivalence, equitable partition, and regular equivalence. Figure 1 shows how these different definitions generate different

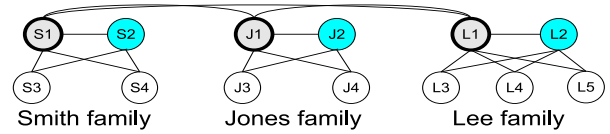


Figure 1: Example Graph for Equivalence Classes.

Equivalence	Neigh. Rule	Non-singleton Classes
Structural	exactly same	$\{S3, S4\}, \{J3, J4\}, \{L3, L4, L5\}$
Automorphic, Exact Color.	same number per class	$\{S1, J1\}, \{S2, J2\}, \{S3, S4, J3, J4\}, \{L3, L4, L5\}$
Regular	same class	$\{S1, J1, L1\}, \{S2, J2, L2\}, \{S3, S4, J3, J4, L3, L4, L5\}$

Table 1: Equivalence Classes for Figure 1

roles from the same network.

Let $G = (V, E)$ be a graph with vertex set $V = \{v_1, \dots, v_n\}$ and edge set E . For any node $v \in V$, let $N(v)$ be the neighbors of v and N_v be the degree of v .

Structural Equivalence: Two actors are *structurally equivalent* if they interact with the *same* set of others [28]. Mathematically, u and v are structurally equivalent if and only if $N(u) = N(v)$. For example, consider the extended family shown in Figure 1. $S1$, $J1$, and $L1$ are siblings, $S2$, $J2$, and $L2$ are spouses, and the remaining nodes are their children. Each family's children, $\{S3, S4\}$, $\{J3, J4\}$, and $\{L3, L4, L5\}$ form a nontrivial equivalence class. However, none of the parents can be grouped together via structural equivalence. This equivalence model is too strict to be useful for simplifying a large network and to discover meaningful roles.

Automorphic Equivalence: Two actors (nodes) u and v are *automorphically equivalent* if there is an automorphism σ of G such that $v = \sigma(u)$ [4]. An automorphism σ of a graph G is a permutation of vertex set V such that for any two nodes u and v , $(u, v) \in E$ iff $(\sigma(u), \sigma(v)) \in E$. In social terms, u and v can swap names, along with possibly some other name swaps, while preserving all the actor-actor relationships. Let $\Gamma(G)$ be the group of all automorphisms of graph G . For any two nodes u and v in G , $u \equiv v$ if $u = \sigma(v)$ for some $\sigma \in \Gamma(G)$. Note that \equiv is an equivalence relation on V ; if $u \equiv v$ we say that u is automorphically equivalent to v . The equivalence classes generated under $\Gamma(G)$ (or \equiv) are called orbits. The equivalence class for vertex $v \in V$ is called the orbit of v , and denoted as $\Delta(v) = \{\sigma(v) \in V, \sigma \in \Gamma(G)\} = \{u | u \equiv v\}$. Each orbit corresponds to a role in the automorphic equivalence. Understanding the importance of automorphic equivalence and applying it to role modeling was a major breakthrough in classical social network research. In our example Figure 1, from the topology alone, we cannot distinguish between the Smith family and the Jones family. The Lee family is distinct, because it has three children instead of two. Therefore, the equivalence classes are $\{S1, J1\}$, $\{S2, J2\}$, $\{S3, S4, J3, J4\}$, $\{L1\}$, $\{L2\}$, and $\{L3, L4, L5\}$. Interestingly, we can observe that automorphically equivalent classes must have equivalent indirect relations as well, such as equivalent in-laws and cousins. However, automorphic equivalence is hard to compute and still very strict.

Exact Coloration (Equitable Partition): An *exact coloration* of graph G assigns a color to each node, such that any two nodes share the same color iff they have the same number of neighbors of each color [11]. Nodes of the same color form an equivalence class. An exact coloration is also referred to as equitable parti-

tion [15] and graph divisor [8] and is often applied in the vertex classification/refinement for canonical labeling of graph isomorphism test [36, 33]. A graph may have several exact colorations; in general we seek the fewest colors. In our running example, the structural equivalence partitioning and the automorphic partitioning offer two different exact colorations. Exact coloration relaxes automorphism by considering only immediate neighborhood equivalence. Two nodes with the same color under an exact coloration may not necessarily be automorphically equivalent, but the graph automorphic equivalence does introduce an exact coloration by assigning a unique color to each orbit. Like automorphic equivalence, exact coloration equivalence provides a recursive aspect to role modeling.

Regular Equivalence (Bisimulation): Two actors are *regularly equivalent* if they interact with the same variety of role classes, where class is recursively defined by regular equivalence [46]. Unlike automorphic equivalence and exact coloration, regular equivalence does not care about the cardinality of neighbor relationships, only whether they are nonzero. For example, using regular equivalence, all three families are now equivalent. There are only three equivalence classes: *sibling* – *parent*{*S1*, *J1*, *L1*}, *spouse* – *parent*{*S2*, *J2*, *L2*}, and *child*. Note that under regular equivalence, any two automorphically equivalent nodes may be partitioned into the same regular equivalence class. In computer science, the regular equivalence is often referred to as the bisimulation, which is widely used in automata and modal logic [32].

3. AXIOMATIC ROLE SIMILARITY

An equivalence relation, however, tells us nothing about non-equivalent items. Using our example, the intuitive and real-world need is for a measure that not only recognizes automorphic equivalence, such as Smith child/spouse/parent to Jones child/spouse/parent, but also tell us that a Lee child/spouse/parent has strong similarity to either a Lee or Smith child/spouse/parent. Over the years, several methods have been developed for addressing various link-based similarity problems (co-citation [39], coupling [21], SimRank [19]). Recently, several researchers have tried to apply these measurements to role modeling [22, 50]. However, none of these encompass the aforementioned automorphic equivalence property and thus are inadequate for measuring role similarity. To deal with this shortcoming and to clarify the problem, we first identify a list of axiomatic properties that all role similarity measures should obey.

DEFINITION 1. (Axiomatic Role Similarity Properties) Given a graph $G = (V, E)$, any $sim(a, b)$ that measures the neighbor-based role similarity between vertices a and b in V should satisfy properties P1 to P5:

- P1) Range: $0 \leq sim(a, b) \leq 1$, for all a and b .
- P2) Symmetry: $sim(a, b) = sim(b, a)$.
- P3) Automorphism confirmation: If $a \equiv b$, $sim(a, b) = 1$.
- P4) Transitive similarity: If $a \equiv b$, $c \equiv d$, then $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$.
- P5) Triangle inequality: $d(a, c) \leq d(a, b) + d(b, c)$, where distance $d(a, c)$ is defined as $1 - sim(a, c)$.

Any node similarity measure satisfying the first four conditions (without triangle inequality) is called an **admissible role similarity measure**. Any node similarity measure satisfying all five conditions is an **admissible role similarity metric**. If the converse of the automorphic confirmation property is also true (if $sim(a, b) = 1$, then $a \equiv b$), then the node similarity measure(metric) is an **ideal role similarity measure(metric)**.

Property 1 describes the standard normalization where 1 means fully similar and 0 means completely dissimilar (i.e., the two neighborhoods have nothing in common). Property 2 indicates that similarity, like distance, must be symmetric. Property 3 expresses our idea that fully similar means automorphically equivalent. Property 4 claims that the similarity between two nodes is equal to the similarity between equivalent members of the first two node's respective equivalence classes. In other words, we can simply define the similarity for the orbits, i.e., $sim(\Delta(u), \Delta(v)) = sim(u, v)$. This guarantees consistency of values at an orbit-level. Property 5 assumes the measure is metric-like, i.e., satisfying the triangle inequality. This is much stronger than transitivity, enforcing an *ordering* of values. Indeed, the only condition which excludes $d(a, b) = 1 - sim(a, b)$ from being a strict distance metric is the automorphic equivalence (it allows the distance between two different nodes to be 0). In addition, note that Property 5 implies Property 4.

LEMMA 1. (Transitive Similarity) For any $a, b \in V$ and $c, d \in V$, if $a \equiv b$ and $c \equiv d$, then $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$.

Proof: From triangle inequality, we have $d(a, c) \leq d(a, b) + d(b, c) \leq d(b, c)$ and $d(b, c) \leq d(b, a) + d(a, c) \leq d(a, c)$ ($d(a, b) = 0$). Thus, $d(a, c) = d(b, c)$. Similarly, $d(a, d) = d(b, d)$, $d(c, a) = d(d, a)$, and $d(d, a) = d(d, b)$. Put together, we have $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$. \square

However, since most similarity measures do not necessarily satisfy the triangle inequality, we explicitly include Property 4 as one of the axiomatic properties. Further, Property 3 is an essential criterion which distinguishes the role similarity measure from other existing measures. As we discussed earlier, the automorphic equivalence can be relaxed to exact coloration or regular equivalence. In this case, we may replace Property 3 accordingly. Our work will focus on the automorphic equivalence though it can handle its generalization as well.

THEOREM 1. (Generalized Transitive Similarity) For any two pairs of nodes $a, b \in V$, $c, d \in V$, if $sim(a, b) = 1$ and $sim(c, d) = 1$, then, their cross similarities are all equal, i.e., $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$.

Proof: From the triangle inequality, we have $d(a, c) \leq d(a, b) + d(b, c) \leq d(b, c)$ and $d(b, c) \leq d(b, a) + d(a, c) \leq d(a, c)$ ($d(a, b) = 0$). Thus, $d(a, c) = d(b, c)$. Similarly, $d(a, d) = d(b, d)$, $d(c, a) = d(d, a)$, and $d(d, a) = d(d, b)$. Put together, we have $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$. \square

Thus, if we partition the nodes into equivalence classes where similarity equals 1, we can simply record the similarity values between equivalent classes. Let $\Delta(x)$ and $\Delta(y)$ be the equivalence classes for node x and y , respectively. Then, we can define $sim(\Delta(x), \Delta(y)) = sim(x, y)$.

3.1 Binary-Valued Role Similarity Measures

THEOREM 2. (Binary Admissibility) Given any equivalence relation that also satisfies automorphism confirmation (P3), its binary indicator function is an admissible similarity metric.

Proof: Binary values satisfy the Range(P1). Any equivalence relation satisfies symmetry (P2) and transitivity (P4), by definition. For triangle inequality (P5), consider all possible cases: Binary values satisfy the Range(P1). Any equivalence relation satisfies symmetry (P2) and transitivity (P4), by definition. For triangle inequality (P5), consider all possible cases:

- Case 1: All in the same class: $0 \leq 0 + 0$
Case 2: All in different classes: $1 \leq 1 + 1$
Case 3: a and c in the same class: $0 \leq 1 + 1$
Case 4: b and one other in the same class: $1 \leq 0 + 1$
□

Note that automorphic equivalence, regular equivalence, and exact coloration all satisfy P3, so they are admissible metrics. In addition, the binary similarity measure introduced by *automorphic equivalence* is an *ideal* role similarity metric. Though these binary similarity measures are admissible, they provide no meaningful information about cross-class similarities, because they set $\text{sim}(\Delta(x), \Delta(y)) = 0$ if $\Delta(x) \neq \Delta(y)$. We would like a real-valued measure that ranks the degree of role similarity.

Before presenting our proposed real-valued role similarity metric for network roles, we first examine some similarity measures proposed in earlier works. We will see that these do not satisfy our required properties.

3.2 SimRank is NOT Admissible

The SimRank [19] similarity between nodes u and v is the average similarity between u 's neighbors and v 's neighbors:

$$SR(u, v) = \frac{(1 - \beta)}{|N(u)||N(v)|} \sum_{x \in N(u)} \sum_{y \in N(v)} SR(x, y), \text{ for } u \neq v,$$

$$SR(v, v) = 1,$$

where β is a decay factor, $0 < \beta < 1$, so that the influence of neighbors decreases with distance. The original SimRank measure is for directed graphs. Here, we focus on its undirected version, though our comments also hold for the directed version. SimRank values can be computed iteratively, with successively iterations approaching a unique solution, much as PageRank [35] does.

THEOREM 3. *SimRank is not an admissible role similarity measure.*

Proof: We give examples where property 3 (automorphic equivalence) does not hold. In Figure 2(a), a and b have the same neighbors. By even the strictest definition (structural equivalence), a and b have the same role. However, since SimRank's *initial* assumption is that there is no similarity among c, d , and e , when it computes the average similarity of a and b 's neighbors, it will never discover their equivalence. Assuming the best case where c, d , and e are structurally equivalent and using the recommended $\beta = 0.15$, $SR(a, b)$ converges to only 0.667. If the neighbors are not equivalent, a to b should still be equivalent, but SimRank gives an even lower value. SimRank has another problem (Figure 2(b)) when there is an odd distance between two nodes. Nodes u and v are automorphically equivalent, but because there are no nodes that are an equal distance from both u and v , $\text{SimRank}(u, v) = 0$!

We note that other variants of SimRank [1, 13, 23, 48, 49, 50] also do not meet the automorphic equivalence property for to similar reasons. More discussion of these variants can be found in the Appendix.

4. ROLESIM: A REAL-VALUED ADMISSIBLE ROLE SIMILARITY

To produce an admissible real-valued role similarity measure, we face two key challenges: First, it is computationally difficult to satisfy the automorphic equivalence property. Though not proven to be NP-complete, the graph automorphism problem has no known polynomial algorithm [14]. Second, all the existing



Figure 2: Problematic configurations for SimRank

real-valued role similarity measures have problems dealing with even simple conditions such as structural equivalence (Subsection 3.2). To meet these challenges, we take the following approach: Given an initial simplistic but admissible role similarity measurement for any pair of nodes in a graph, refine the measurement by expressing the similarity in terms of neighboring values, while maintaining the automorphic and structural equivalence properties. In the following, we formally introduce RoleSim, the first admissible real-valued role similarity measure (metric) and its associated properties.

4.1 RoleSim Definition

Given a graph $G = (V, E)$, the RoleSim measure realizes the recursive node structural similarity principle “two nodes are similar if they relate to similar objects” as follows.

DEFINITION 2. (RoleSim metric) *Given two vertices u and v , where $N(u)$ and $N(v)$ denote their respective neighborhoods and N_u and N_v denote their respective degrees, then $\text{RoleSim}(u, v) =$*

$$(1 - \beta) \max_{M(u, v)} \frac{\sum_{(x, y) \in M(u, v)} \text{RoleSim}(x, y)}{N_u + N_v - |M(u, v)|} + \beta \quad (1)$$

where $x \in N(u)$, $y \in N(v)$, and $M(u, v)$ is a **matching** between $N(u)$ and $N(v)$, i.e., $M(u, v) = \{(x, y) | x \in N(u), y \in N(v), \text{ and no other } (x', y') \in M(u, v), \text{ s.t. } x = x' \text{ or } y = y'\}$. The parameter β is a decay factor, $0 < \beta < 1$.

The decay factor, similar to the one used in PageRank [35], both dampens the recursive effect and guarantees a minimal RoleSim score of β . We will sometimes abbreviate $\text{RoleSim}(u, v)$ as $R(u, v)$. \mathbf{R} refers to the entire matrix of values. Figure 3 illustrates the matching process. The (x, y) grid is the subset of the RoleSim matrix of values corresponding to the pairings of neighbors of these two vertices. A matching selects one cell per row and column. If the number of rows differs from the number of columns, then the matching size is limited to $|M(u, v)| = \min(N_u, N_v)$. A maximal matching is a matching where the total value of selected cells is maximum. In contrast, SimRank computes the average of every cell in the neighbor grid.

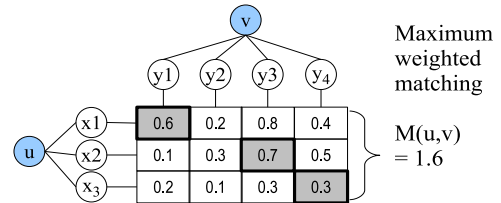


Figure 3: RoleSim(a,b) based on similarity of their neighbors

4.1.1 Relation to Jaccard Coefficient

RoleSim is built on top of a natural generalization of the Jaccard coefficient, which measures the similarity between two sets A and B as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. The Jaccard coefficient has been used previously to measure node-node similarity based on their neighborhood commonality [13]. In our generalization, however, sets A and B do not necessarily share any common element; instead, there is a matching M between *similar* elements in A and B , i.e., $(a, b) \in M, a \in A, b \in B$. Let $r(a, b) \in [0, 1]$ record the similarity between a and b .

DEFINITION 3. (Generalized Jaccard Coefficient) *The generalized Jaccard coefficient measures the similarity between two sets A and B under matching M , defined as*

$$J(A, B|M) = \frac{\sum_{(a,b) \in M} r(a, b)}{|A| + |B| - |M|} \quad (2)$$

The original Jaccard coefficient is a special case which uses the following matching M : Let $r(x, y) = 1$ if $x = y$; otherwise 0. Then define $M = \{r(x, x) | x \in A, x \in B\}$. Thus, the generalized Jaccard coefficient $J(A, B|M)$ reduces to $J(A, B)$. Comparing Eq. (1) and (2), we see that the heart of $RoleSim(u, v)$ is equivalent to the maximum of the generalized Jaccard coefficient between $N(u)$ and $N(v)$, among all matchings $M(u, v)$. Then, $RoleSim(u, v) =$

$$(1 - \beta) \max_{M(u, v)} J(N(u), N(v) | M(u, v)) + \beta \quad (3)$$

4.1.2 Relation to Weighted Matching

The definition and significance of the RoleSim for any node pair (u, v) is closely related to *maximal weighted matching*. For any nodes u and v in graph G , define a weighted bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$, with each edge $(x, y) \in N(u) \times N(v)$ having weight $RoleSim(x, y)$. Let the total weight of neighbor matching $M(u, v)$ between u and v be $w(M(u, v)) = \sum_{(x, y) \in M(u, v)} RoleSim(x, y)$. Let \mathcal{M} be the maximal weighted matching for $(N(u) \cup N(v), N(u) \times N(v))$. It is clear that

$$w(\mathcal{M}) = \max_{M(u, v)} w(M(u, v)). \quad (4)$$

Using this, we can represent $RoleSim(u, v)$ in terms of maximal weighted matching \mathcal{M} . In Figure 3, the shaded cells represent the maximal matching: $0.7 + 0.6 + 0.3 = 1.6$.

THEOREM 4. (Maximal Weighted Matching) *The RoleSim between nodes u and v corresponds linearly to the maximal weighted matching \mathcal{M} for the bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$, with each edge $(x, y) \in N(u) \times N(v)$ having the weight $RoleSim(x, y)$:*

$$RoleSim(u, v) = (1 - \beta) \frac{w(\mathcal{M})}{\max(N_u, N_v)} + \beta \quad (5)$$

Proof: We need to show that Equations (1) and (5) are equivalent. Without loss of generality, let $N_u \geq N_v$. First, we show that the cardinality of the maximal weighted matching $|\mathcal{M}| = \min(N_u, N_v) = N_v$. It cannot be greater, because there are insufficient elements in N_v . It cannot be smaller, because if it were, there must exist an available edge between an uncovered node in N_u with one in N_v . Adding this edge would increase the matching (every edge has weight $\geq \beta$). If $|\mathcal{M}| = \min(N_u, N_v)$, it follows that $N_u + N_v - |\mathcal{M}| = \max(N_u, N_v)$. Thus, the denominators in Equations (1) and (5) are constant and identical. It is then a trivial observation that the numerators are in fact the same. Therefore, the maximal value for the entire Equation (1) is the same as the value in (5). \square

Theorem 4 not only shows the key equilibrium for the role similarities RoleSim between pairs of nodes in a graph G , but shows that each iteration can be computed using existing maximal matching algorithms.

4.2 RoleSim Computation

RoleSim values can be computed iteratively and are guaranteed to converge, just as in PageRank and SimRank. First we outline the procedure. In the next section, we prove that the calculated values comprise an admissible role similarity metric.

Step 1: Let the initial matrix of RoleSim scores \mathbf{R}^0 be any set of admissible scores between any pair of nodes in G .

Step 2: Compute the k^{th} iteration \mathbf{R}^k scores from the $(k - 1)^{th}$ iteration's values, \mathbf{R}^{k-1} . Specifically, for any nodes u and v ,

$$\mathbf{R}^k(u, v) = (1 - \beta) \max_{M(u, v)} \frac{\sum_{(x, y) \in M(u, v)} \mathbf{R}^{k-1}(x, y)}{N_u + N_v - |M(u, v)|} + \beta \quad (6)$$

Based on Theorem 4, we compute Equation (6) by finding the maximal weighted matching in the weighted bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$ with each edge $(x, y) \in N(u) \times N(v)$ having weight $\mathbf{R}^{k-1}(x, y)$.

Step 3: Repeat Step 2 until \mathbf{R} values converge for each pair of nodes in G .

THEOREM 5. (Convergence) *For any admissible set of RoleSim scores $RoleSim^0$, the iterative computational procedure for RoleSim converges, i.e., for any (u, v) pair,*

$$\lim_{k \rightarrow \infty} RoleSim^k(u, v) = RoleSim(u, v) \quad (7)$$

This can be proven by showing that the maximum absolute difference between any $\mathbf{R}^k(u, v)$ and $\mathbf{R}^{k+1}(u, v)$ is monotonically decreasing. The proof is in the Appendix.

Unlike PageRank and SimRank which converge to values independent of the initialization, the convergent RoleSim score is sensitive to the initialization. That is, different initial values may generate different final RoleSim values. Rather than being a disadvantage, this is actually the key to coping with the graph automorphism complexity, by allowing the ranking to utilize prior knowledge (the equivalence relationship) of the network topological structure.

4.3 Admissibility of RoleSim

Here, we present one of the key contributions of this paper: the axiomatic admissibility of RoleSim. If the initial computation is admissible, and because the iterative computation of Equation (5) maintains admissibility (i.e., is an invariant transform of the axiomatic properties), then the final measure is admissible.

THEOREM 6. (Invariant Transformation) *If the k^{th} iteration $RoleSim^k$ is an admissible role similarity metric, then so is $RoleSim^{k+1}$.*

Properties 1 (Range) and 2 (Symmetry) are trivially invariant, so we will focus on Properties 3 (Automorphic Equivalence), 4 (Transitive Similarity), and 5 (Triangle Inequality).

LEMMA 2. (Automorphism Confirmation Invariance) *If the k^{th} iteration $RoleSim^k$ satisfies Axiom 3 (Automorphism Confirmation), then so does $RoleSim^{k+1}$.*

Proof: For nodes $u \equiv v$, there is a permutation σ of vertex set V , such that $\sigma(u) = v$, and any edge $(u, x) \in E$ iff

$(v, \sigma(x)) \in E$. This indicates that σ provides a one-to-one equivalence between nodes in $N(u)$ and $N(v)$. Also, u and v have the same number of neighbors, i.e., $N_u = N_v$. So, it is clear that the maximal weighted matching \mathcal{M} in the bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$ selects $N_u = N_v$ pairs of weight 1 each. Thus, $\text{RoleSim}^{k+1}(u, v) = (1 - \beta) \frac{w(\mathcal{M})}{\max(N_u, N_v)} + \beta = 1$. \square

LEMMA 3. (Transitive Similarity Invariance) *If the k^{th} iteration RoleSim^k satisfies Axiom 4 (Transitive Similarity), then so does RoleSim^{k+1} .*

Proof: We know for any $a \equiv b, c \equiv d$, $\text{RoleSim}^k(a, c) = \text{RoleSim}^k(b, d)$. Denote the maximal weighted matching between $N(a)$ and $N(c)$ as \mathcal{M} . Since there is a one-to-one equivalence correspondence σ between $N(a)$ and $N(b)$ and a one-to-one equivalence correspondence σ' between $N(c)$ and $N(d)$, we can construct a matching \mathcal{M}' between $N(b)$ and $N(d)$ as follows: $\mathcal{M}' = \{(\sigma(x), \sigma'(y)) | (x, y) \in \mathcal{M}\}$. Since the transitive similarity property holds for RoleSim^k , we have $\text{RoleSim}^k(x, y) = \text{RoleSim}^k(\sigma(x), \sigma'(y))$. Thus, $w(\mathcal{M}') = w(\mathcal{M})$, and

$$(1 - \beta) \frac{w(\mathcal{M})}{\max(N_a, N_c)} + \beta = (1 - \beta) \frac{w(\mathcal{M}')}{\max(N_b, N_d)} + \beta$$

$$\text{RoleSim}^{k+1}(a, c) = \text{RoleSim}^{k+1}(b, d).$$

\square

LEMMA 4. (Triangle Inequality Invariance) *If the k^{th} iteration RoleSim^k satisfies Axiom 5 (Triangle Inequality), then so does RoleSim^{k+1} .*

Proof: For iteration k , for any nodes a, b , and c , $d^k(a, c) \leq d^k(a, b) + d^k(b, c)$, where $d^k(a, b) = 1 - \text{RoleSim}^k(a, b)$. We must prove that this inequality still holds for the next iteration: $d^{k+1}(a, c) \leq d^{k+1}(a, b) + d^{k+1}(b, c)$.

Observation: if there is a matching M between $N(a)$ and $N(c)$ which satisfies $1 - ((1 - \beta) \frac{w(M)}{N_c} + \beta) \leq d^{k+1}(a, b) + d^{k+1}(b, c)$, then $d^{k+1}(a, c) \leq d^{k+1}(a, b) + d^{k+1}(b, c)$. This is because $\frac{w(M)}{N_c} \leq \frac{w(\mathcal{M})}{N_c}$, where \mathcal{M} is the maximal weighted matching between $N(a)$ and $N(c)$, and thus, $1 - ((1 - \beta) \frac{w(M)}{N_c} + \beta) \geq 1 - ((1 - \beta) \frac{w(\mathcal{M})}{N_c} + \beta) = d^{k+1}(a, c)$.

We break down the proof into three cases:

Case 1. ($N_b \leq N_a \leq N_c$), Case 2. ($N_a \leq N_b \leq N_c$), and Case 3. ($N_a \leq N_c \leq N_b$).

Case 1 ($N_b \leq N_a \leq N_c$): Since N_b is smallest, $|\mathcal{M}(a, b)| = |\mathcal{M}(b, c)| = N_b$. Define matching M between $N(a)$ and $N(c)$ as $M = \{(x, z) | (x, y) \in \mathcal{M}(a, b) \wedge (y, z) \in \mathcal{M}(b, c)\}$. Then

using our observation above:

$$\begin{aligned} & d^{k+1}(a, b) + d^{k+1}(b, c) - (1 - (1 - \beta) \frac{w(M)}{N_c} - \beta) \\ &= (1 - \beta) \left[-\frac{w(\mathcal{M}(a, b))}{N_a} - \frac{w(\mathcal{M}(b, c))}{N_c} + \frac{w(M)}{N_c} \right] + 1 - \beta \\ &= (1 - \beta) \left[\frac{N_b - w(\mathcal{M}(a, b))}{N_a} - \frac{N_b}{N_a} + \frac{N_b - w(\mathcal{M}(b, c))}{N_c} \right. \\ &\quad \left. - \frac{N_b}{N_c} - \frac{N_b - w(M)}{N_c} + \frac{N_b}{N_c} \right] + 1 - \beta \\ &\geq (1 - \beta) \left[1 - \frac{N_b}{N_a} + \frac{\sum_{(x, y) \in \mathcal{M}(a, b)} (1 - R^k(x, y))}{N_c} \right. \\ &\quad \left. + \frac{\sum_{(y, z) \in \mathcal{M}(b, c)} (1 - R^k(y, z))}{N_c} - \frac{\sum_{(x, z) \in M} (1 - R^k(x, z))}{N_c} \right] \\ &\geq (1 - \beta) \left[\frac{\sum_{(x, y, z)} (d^k(x, y) + d^k(y, z) - d^k(x, z))}{N_c} \right] \geq 0 \end{aligned}$$

where $(x, y) \in \mathcal{M}(a, b)$, $(y, z) \in \mathcal{M}(b, c)$, $(x, z) \in M$

Cases 2 and 3 can be proven by a similar technique; the details are in the Appendix.

By combining the admissible initial configurations given in Sec 4.4 with Theorem 6 on invariance, we have shown that the iterative RoleSim computation generates a real-valued, admissible role similarity measure.

THEOREM 7. (Admissibility) *If the initial RoleSim^0 is an admissible role similarity measure, then at each k -th iteration, RoleSim^k is also admissible. When RoleSim computation converges, the final measure $\lim_{k \rightarrow \infty} \text{RoleSim}^k$ is admissible.*

4.4 Initialization

According to Theorem 7, an initial admissible RoleSim measurement $\mathbf{R}^0 = I(\cdot)$ is needed to generate the desired real-valued role similarity ranking. What initial admissible measures or prior knowledge should we use? We consider three schemes:

1. **ALL-1**: $I(u, v) = 1$ for all u, v .
2. **Degree-Binary (DB)**: If two nodes have the same degree ($N_u = N_v$), then $I(u, v) = 1$; otherwise, 0.
3. **Degree-Ratio (DR)**: $I(u, v) = (1 - \beta) \frac{\min(N_u, N_v)}{\max(N_u, N_v)} + \beta$.

These schemes come from the following observation: *nodes that are automorphically equivalent have the same degree*. Basically, equal degree is a necessary but not sufficient condition for automorphism. This observation is key to RoleSim: degree affects both the size of a maximal matching set and the denominator of the Jaccard Coefficient.

THEOREM 8. (Admissible Initialization) *ALL-1, Degree-Binary, and Degree-Ratio are all admissible role similarity measures. Moreover, Degree-Binary and ALL-1 are admissible role similarity metrics.*

Proof: It is easy to see that ALL-1 degenerately satisfies all the axioms of a role similarity metric. We focus on the two degree-based schemes. Clearly, they satisfy Range(P1) and Symmetry(P2). If $N_u = N_v$, then $I(u, v) = 1$, so they both satisfy Automorphism Confirmation (P3). For transitive similarity (P4), we only need to show that $I(u, v)$ depends only on class membership (Theorem 1). For these schemes, class is defined by degree, and the measurement clearly depends only on degree. Finally, because Degree-Binary and ALL-1 are binary indicators of equivalence, Theorem 2 states that they are metrics. \square

Note that SimRank's initialization ($\text{SimRank}^0(u, v) = 1$ iff $u = v$) is NOT admissible, because it does exactly the wrong thing: setting the initial value of any potentially equivalent nodes to 0. SimRank iterations try to build up from zero. However, due to its problems with structural equivalence and odd-length paths that we noted, SimRank will never increase the value enough to discover equivalent pairs that were neglected at the start.

In addition, we make the following interesting observations on the different initialization schemes.

LEMMA 5. *Let $\mathbf{R}^1(\text{ALL} - 1)$ be the matrix of RoleSim values at the first iteration after $\mathbf{R}^0 = \mathbf{1}$ (All-1 initialization). Let $\mathbf{R}^0(\text{DR})$ be the matrix of RoleSim initialized by the Degree-Ratio (DR) scheme. Then, $\mathbf{R}^1(\text{ALL} - 1) = \mathbf{R}^0(\text{DR})$.*

This lemma can be easily derived by following the definition of RoleSim formula. Basically, the Degree-Ratio (DR) is exactly equal to the RoleSim state one iteration after ALL-1 initialization. Thus, ALL-1 and DR generate the same final results. The simple formula for DR is much faster than neighbor matching, so DR is essentially one iteration faster. On the other hand, we may consider the simple ALL-1 scheme to be sufficient, since it works as well as the more sophisticated DR. Especially, after the simple initialization, RoleSim's maximal matching process automatically discriminates between nodes of different degree and continues to learn differences among neighbors as it iterates. Also, both ALL-1 and DR initialization have the following convergence property:

THEOREM 9. (Monotone Convergence) *If ALL-1 initialization is used, each RoleSim value is monotonically decreasing (or non-increasing): $\mathbf{R}^{k+1}(u, v) \leq \mathbf{R}^k(u, v)$ for all k .*

Proof: At any iteration, the RoleSim value for any (u, v) is the maximal matching of its neighbors. The value can increase only if some neighbor matchings increase. If no value increased in the previous iteration, then no value can increase in the current iteration. In the first iteration after ALL-1, clearly no value increases. Therefore, no value ever increases. \square

Indeed, this monotone convergence property can be generalized into the following format: *if $\mathbf{R}^1 \leq \mathbf{R}^0$ (for any (u, v) pair, $\mathbf{R}^1(u, v) \leq \mathbf{R}^0(u, v)$), then we have $\mathbf{R}^{k+1} \leq \mathbf{R}^k$.* Note that the Degree-Binary (DB) initialization scheme does not have this property. In our experiments, we will further empirically study these initialization schemes.

4.5 Computational Complexity

Given n nodes, we have $O(n^2)$ node-pair similarity values to update for each iteration. For each node-pair, we must perform a maximal weighted matching. For weighted bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$, the fastest algorithm based on augmenting paths (Hungarian method) can compute the maximal weighted matching in $O(x \log x + y)$, where $x = |N(u) \cup N(v)|$ and $y = |N(u)| \times |N(v)|$.

A fast greedy algorithm offers a $\frac{1}{2}$ -approximation of the globally optimal matching in $O(y \log y)$ time [2]. If an equivalence matching exists (i.e., $w(\mathcal{M}) = \max(N_u, N_v)$), the greedy method will find it. This is important, because it means that a greedy RoleSim computation still generates an admissible measure. Using greedy neighbor matching, the overall time complexity of RoleSim is $O(kn^2 d')$, where k is the number of iterations and d' is the average of $y \log y$ over all vertex-pair bipartite graphs in G . The space complexity is $O(n^2)$.

5. ICEBERG ROLESIM COMPUTATION

Node similarity ranking in general is computationally expensive because we need to compute the similarity for $\binom{n}{2} = O(n^2)$ node-pairs. A graph with 100,000 nodes needs about 40GB memory to simply maintain the similarity values, assuming 8 bytes per value. Indeed, this is a major problem for almost all node similarity ranking algorithms. However, in most applications, we are interested only in the *highest* similarity pairs, which typically compose only a very small fraction of all pairs. Thus, in order to improve the scalability of RoleSim, we ask the following question: *Can we identify the high-similarity pairs without computing all pair similarities?* Formally, we consider the following question:

DEFINITION 4. (Iceberg RoleSim) *Given a threshold θ , the Iceberg RoleSim problem is to discover all (u, v) pairs for which $\text{RoleSim}(u, v) \geq \theta$ and then approximate their RoleSim scores.*

The goal is to identify and compute those high-similarity pairs without materializing the majority of the low similarity pairs. To solve *Iceberg RoleSim*, we consider a two-step approach: 1) use pruning rules to rule out pairs whose similarity score must be less than θ ; and 2) apply RoleSim iterative computation to the remaining candidate pairs. Since RoleSim computation must match all neighbor-pairs ($N(u) \times N(v)$) of a candidate pair (u, v) , we have to handle neighbor-pairs (such as x, y) which are not themselves candidate pairs. Here, we employ upper and lower bounds for estimating RoleSim values for the non-candidate pairs.

Upper and Lower Bound for RoleSim:

LEMMA 6. *Given nodes u, v and without loss of generality, $N_u \geq N_v$, if $N_v \leq \theta N_u$, then similarity $R(u, v) \leq (1 - \beta)\theta + \beta$.*

Proof: $R(u, v) = (1 - \beta) \frac{w(\mathcal{M})}{N_u} + \beta \leq (1 - \beta) \frac{N_v}{N_u} + \beta \square$

Given this, assuming $N_u \geq N_v$, since matching $0 \leq w(\mathcal{M}) \leq N_v$, then $R(u, v)$ is in the range $[\beta, (1 - \beta) \frac{N_v}{N_u} + \beta]$. Furthermore, to facilitate our discussion, we further define $\theta' = (\theta - \beta) / (1 - \beta)$. Now, we introduce the following *pruning rules* to filter out those pairs whose RoleSim cannot be greater than or equal to threshold θ , without knowing their exact *RoleSim* scores (Without loss of generality, let $N_u \geq N_v$):

1. If $N_v < \theta' N_u$, then $R(u, v) < \theta$
2. If maximal matching weight $w(\mathcal{M}) < \theta' N_u$, then $R(u, v) < \theta$
3. Assume neighbor lists $N(u)$ and $N(v)$ are sorted by degree, with d_1^u and d_1^v being the first items. The maximum possible similarity of this pair is $m_{11} = (1 - \beta) \frac{\min(d_1^u, d_1^v)}{\max(d_1^u, d_1^v)} + \beta$. If the shorter list has the smaller degree ($d_1^v \leq d_1^u$), and if $m_{11} + N_v - 1 < \theta' N_u$, then $R(u, v) < \theta$.

Rule 1 is just a restatement of Lemma 6. Rule 2 is based on the upper bound of RoleSim value. Rule 3 requires more explanation: continuing from Rule 2, we begin to consider all the pairings of neighbors. Because N_v is the shorter list, every member must contribute to the final matching. Either m_{11} will be in the matching or not. If it is, then an upper bound for \mathcal{M} is if every remaining pair has weight 1, yielding $m_{11} + (N_v - 1)$. Additionally, because the lists are sorted, $d_1^u / d_1^v \geq d_x^u / d_x^v$, for $x > 1$. So, if m_{11} is too small to satisfy Rule 2, then all pairings using d_1^v are too small. This rule allows us to shortcircuit the full neighbor matching

We now outline our approach, which is formalized in Algorithm 1. To generate the initial iceberg hash map, we sort nodes by degree (line 3) and sort each node's list of neighbors, by degree (lines 4 to 6). The first sort allows us to consider only those node-pairs that are sufficiently similar in degree (line 8, pruning rule

1). We compute the estimated similarity for the first pair of neighbors. Note that this estimation formula is the same as Degree-Ratio initialization. If this weight is below the limit defined in Rule 3, we terminate this pair’s candidacy and move on (lines 9 to 12). Otherwise, compute the remainder of neighbor-pair initial similarities, and perform a maximal matching. If the matching weight exceeds the θ' minimum bound (Rule 2), then this node-pair and its similarity are inserted into the hash table (lines 13 to 16). After iterating through all qualified node-pairs, we have our full hash table. We now perform RoleSim iterations, but only on members of the table, which is orders of magnitude smaller than a complete similarity matrix. When a non-candidate pair’s value is needed (as a neighbor-pair of a candidate pair), we apply the following estimate based on its lower and upper bound (assuming $N_u \geq N_v$):

$$\tilde{R}(u, v) = \alpha(1 - \beta) \frac{N_v}{N_u} + \beta, \text{ where } 0 \leq \alpha \leq 1.$$

In the experimental evaluation, we will empirically study the effect of α on the estimation accuracy.

Algorithm 1 IcebergRoleSim($G(V, E), \theta, \beta, \alpha$)

```

1:  $H \leftarrow$  empty hash table indexed by node-pair ID  $(u, v)$ ;
2:  $d(v) \leftarrow$  degree of  $v$ ;
3: Sort vertices  $V$  by degree;
4: for each  $v \in V$  do
5:    $D^v = \{d_1^v, d_2^v, \dots, d_{d(v)}^v\} \leftarrow$  degrees of neighbors of  $v$ ,
     sorted by increasing order;
6: end for
7: for each  $u \in V$  do
8:   for each  $v \in V, \theta' d(u) \leq d(v) \leq d(u)$  (Rule 1) do
9:      $m_{11} \leftarrow (1 - \beta) \frac{\min(d_1^u, d_1^v)}{\max(d_1^u, d_1^v)} + \beta$ ;
10:    if  $d_1^v \leq d_1^u$  and  $N_v - 1 + M_{11} < \theta' N_u$  then
11:      Skip to the next  $v$ ; (Rule 3)
12:    end if
13:    Compute maximal matching weight  $w(\mathcal{M})$ ;
14:    if  $w(\mathcal{M}) \geq \theta' d(u)$  (Rule 2) then
15:      Insert  $H(u, v) \leftarrow (1 - \beta)w(\mathcal{M})/d(u) + \beta$ ;
16:    end if
17:  end for
18: end for
19: Perform iterative RoleSim on  $H$ . For neighbor pairs  $\notin H$ ,
    use  $\tilde{R}(x, y) = \alpha(1 - \beta)N_x/N_y + \beta$ 

```

6. EXPERIMENTAL EVALUATION

In this section we experimentally investigate the ranking ability and performance of the RoleSim algorithm for computing role similarity metric values. We compare RoleSim to several state-of-the-art node similarity algorithms, analyze the effect of different initialization schemes, and measure the scalability of Iceberg RoleSim. Specifically, we focus on the following questions:

1. How do different initialization schemes perform in terms of their final RoleSim score and computational efficiency?
2. Do node-pairs with high RoleSim scores actually have similar network roles? For any two nodes known to have similar network roles, do they receive high role similarity scores?
3. How much less memory and time does Iceberg RoleSim use, and how closely does its rankings match standard RoleSim’s?

Relative to All-1 Initialization	Degree-Binary			Degree- Ratio
	Min	Avg.	Max	
Diff. in percentile rank	0.14%	0.38%	11.17%	none
Pearson correl. coeff.	0.9994	0.9998	0.9999	1
Relative execution time	0.32	0.52	0.80	≈ 0.9
Relative # iterations	0.38	0.58	0.88	1 fewer

Table 2: Comparison of Initialization Methods

Clearly, the ideal validation study requires an explicit role model and role similarity measure, which often do not exist. In the following study, we utilize a well-known role-related random graph model and external measures of real datasets which provide strong role indication for these evaluations.

We set $\beta = 0.1$ for both RoleSim and SimRank, defining convergence to be when values change by less than 1% of their previous values. We ran several RoleSim tests with both exact matching and greedy matching. The results were nearly identical ($> 90\%$ of cells have no difference; maximum difference was small), so we focus on greedy matching from here on. We implemented the algorithms in C++ and ran all large tests on a 2.0GHz Linux machine with dual-core Opteron CPU and 4.0GB RAM.

For our tests, we use three types of graphs:

- **BL**: the probabilistic block-model [44], where each block is generally considered to be corresponding to a role [47]. Here, nodes are partitioned into blocks. Each node in block i has probability p_{ij} of linking to each node in block j . Thus, the underlying block-model may serve as the ground-truth for testing role similarity.
- **SF**: Large Scale-Free random graphs² are used for testing scalability of the Iceberg RoleSim computation.
- Real-world networks, with a measureable feature similar to social role, are used for validating RoleSim performance.

6.1 Comparing Initialization

In Section 4.4 we discussed that Degree-Ratio initialization generates the same results as ALL-1 by shortcutting the first iteration. This reduces the computation time by roughly 10%. Now we ask: Does Degree-Binary initialization (DB, binary indicator which equals 1 when degrees $N_u = N_v$) give similar results, quickly?

We ran RoleSim using both ALL-1 and DB on 12 graphs, some scale-free and some block-model, having 500 to 10,000 nodes, and edge densities from 1 to 10. We then converted values to percentile ranking, where 100% means the highest value and 50% is the median value. Test results are summarized in Table 2. The high correlation coefficient means the rankings are virtually identical, so the rankings are not very sensitive to the initialization method. Moreover, DB took 20% from 68% less time to converge. Overall, DB seems to be the preferred initialization scheme in terms of computational efficiency. Thus, we adopt it for the rest of the experiments.

6.2 General Role Detection

How well does RoleSim discover roles in complex graphs? Specifically, given a ground truth knowledge of roles, do nodes having similar roles have high scores? To answer this question, we generated probabilistic block-model graphs, where blocks behave like "noisy" roles, due to sampling variance. We generated graphs with $N = 1000$ nodes and either 3 or 5 blocks. We varied the edge density $\frac{|E|}{|V|}$, with higher densities for graphs with more blocks. The size of each block and the p_{ij} values were randomized; we

²<http://pywebgraph.sourceforge.net/>

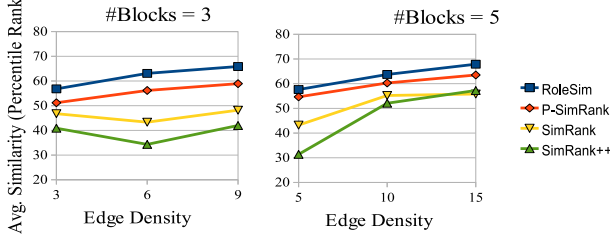


Figure 4: Avg. similarity ranking for nodes in the same block

generated 3 random instances for each graph class. We compared RoleSim to the state-of-the-art SimRank, SimRank++ [1], and P-SimRank [13].

For each measure and trial, we ranked the similarity scores. This serves to normalize the scoring among the four measures. Then, for each graph, we computed the average ranking of all pairs of nodes within the same block. We then averaged the three trials for each graph class.

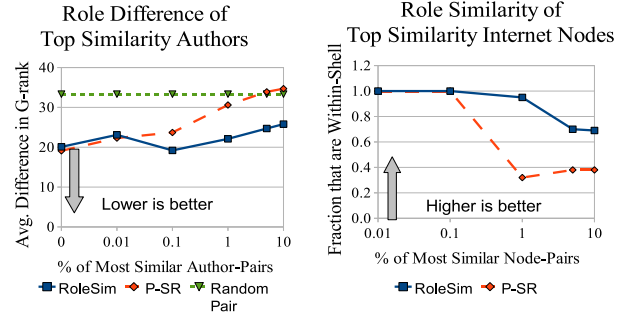
Our results (Figure 4) show that RoleSim outperforms all other algorithms across all the tested conditions. None of the algorithms score perfectly, due to the inherent edge distribution variance of the probabilistic model. P-SimRank is better than SimRank, perhaps because it uses Jaccard Coefficient weighting, a step towards our RoleSim approach. Accuracy takes time. SimRank and SimRank++ run at the same speed. P-SimRank is about 1.5 to 2 times slower, and standard RoleSim is about twice as slow as SimRank.

6.3 Real Dataset: Co-author Network

We applied RoleSim and the best alternative measure, P-SimRank, to a real-world network having an external role measure. Our first dataset [41] is a co-author network of 2000 database researchers. Two authors are linked if they co-authored a paper from 2003 to 2008. We pruned the network to the largest connected component (1543 nodes, 15483 edges). An author's role depends recursively on the number of connections to other authors, and the roles of those others. Hence, it measures collaboration. We use the G-index as a proxy measure for co-author role (H-index provides similar results and thus is omitted here). The G-index measures the influence of a scientific author's publications, its value being the largest integer G such that the G most cited publications have at least G^2 citations. While G-index and co-author role are not precisely the same, G-index score is influenced strongly by the underlying role. High impact authors tend to be highly connected, especially with other high impact authors. If a paper is highly cited, this boosts the score of every co-author. Thus, we expect that if two authors have similar G-index scores, their node-pair is likely to have a high role similarity value. To normalize RoleSim, P-SimRank, and G-index values, we converted each raw value to a percentile rank.

Figure 5(a) addresses our second validation question (high rank \rightarrow similar roles?). For the top ranked 0.01% of author-pairs, their difference in G-index ranking is about 20 points, for both RoleSim and P-SimRank, well below the random-pair value of 33. A below-average difference confirms that the authors are relatively similar. However, as we expand the search towards 10%, RoleSim continues to detect authors with similar authorship performance, while P-SimRank converges to random scoring.

To validate *role* \rightarrow *rank* performance, we binned the authors into 10 roles based on G-index value (bottom 10%, next 10%, etc.). For every pair of authors within the same role decile, we looked up role similarity percentile rank and computed an average



(a) Top Coauthors

(b) Top Internet nodes

Figure 5: Similarity of Nodes for Top Ranked Node-Pairs

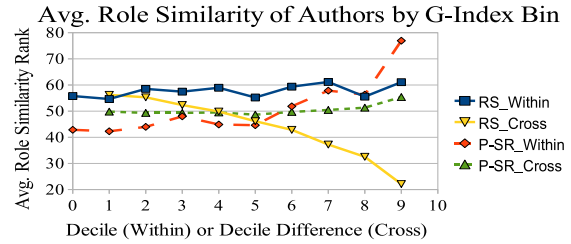


Figure 6: Similarity of Authors Binned by K-index

per bin. We also computed averages for pairs of authors not in the same bin (dissimilar roles). Figure 6 shows our results. The average within-bin RoleSim value is consistently between 55% and 60%, better than the random-pair score of 50, and independent of whether the G-index is high or low. It performs equally well for all roles. P-SimRank within-bin scores (dashed line), however, are inconsistent. Performance of P-SimRank is worse than random for low G-scores, perhaps due to low density of links in the network. For the cross-bin data, the X-axis is the difference in decile bins for the two authors in a pair. The falling line of RoleSim indicates that role similarity correctly decreases as G-index scores become less similar. For P-SimRank, however, the cross-bin scores (dashed line) hover around 50, equivalent to random scoring.

6.4 Real Dataset: Internet Network

Our second dataset is a snapshot of the Internet at the level of autonomous systems (22963 nodes and 48436 edges), as generated by [34]. Several studies have confirmed that the Internet is hierarchically organized, with a densely connected core and stubs (singly-connected nodes) at the periphery [43, 7]. A node's position within the network (proximity to the core) and its relation to others (such as density of connections) affects its efficiency for routing and its robustness. Inspired by [7], we use K -shells to delineate roles.

The K -core of a graph is the induced subgraph where every node connects to at least K other nodes in the subgraph. If $K' > K$, then the K' -core must be an induced subgraph of the K -core. The K -shell is defined as the 'ring' of nodes that are included in a graph's $(K - 1)$ -core but not its K -core. In other words, we can decompose a graph into a set of nested rings, becoming denser as we move inward.

Using K -shells as our roles, we perform tests and analyses similar to those of the coauthor network. In Figure 5(b) we see that

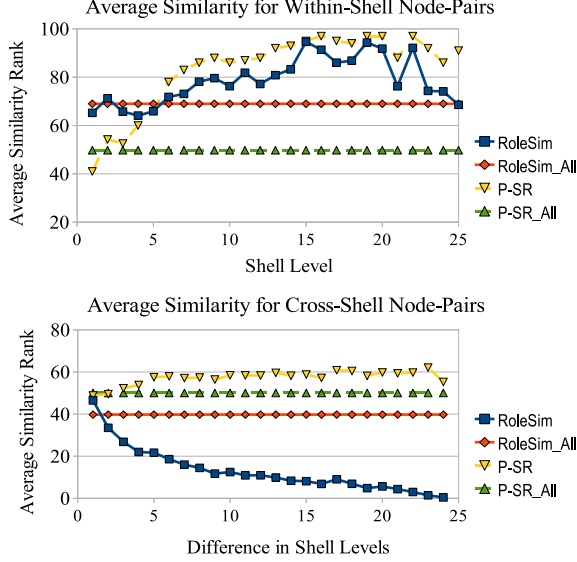


Figure 7: Similarity of Authors Grouped By K-Shell

both measures do well for the top 0.1%, but P-SimRank’s falters significantly when the range is expanded to the top 1%.

Next, we treat K -shells the same way that we treated G-index decile bins in the previous test. See Figure 7. Unlike decile bins, the shells do not have equal sizes. K -shells 1, 2, and 3 together contain 92% of all nodes. To clarify how these three shells dominate, we also show horizontal lines representing the combined weighted average rank of all within-shell comparisons. RoleSim’s within-shell values are consistently high, averaging 70%. Conversely, P-SimRank finds strong above-average similarity for the small high- K shells, but nearly random similarity for shells 1 to 3, pulling its overall performance down to 50%.

In cross-shell analysis, RoleSim is able to distinguish different shells very well: RoleSim approaches zero as shell difference approaches maximum. On the other hand, P-SimRank shows almost no correlation to shell difference. Many of its scores are above-average when they should be below-average (dissimilar). On the whole, it seems that P-SimRank is not detecting role, but something related to connectedness and density.

In all these experiments, we can see that RoleSim provides positive answer to the role similarity ranking: 1) node-pairs with similar roles have higher RoleSim ranking than node-pairs with dissimilar roles, and 2) high RoleSim ranking indicates that nodes have similar roles. P-SimRank scores, however, do not correlate with network role similarity.

6.5 Performance of Iceberg RoleSim

In this experiment, we study how Iceberg RoleSim performs in terms of reducing computational time and storage, and its accuracy at approximating the RoleSim score for high similar node-pairs. Here, we generated 12 scale-free graphs with up to 100K nodes and edge densities of 1, 2, and 5. We compared standard RoleSim to Iceberg RoleSim, with θ values of 0.8 and 0.9. The parameter α , which is the weighting for estimated non-stored values, is set to midpoint 0.5. For the scale-free graphs, the relative scale of the iceberg compared to the full similarity matrix depends on θ and edge density, but it is almost independent of the number of nodes. Table 3 shows that the icebergs’ hash tables are only 0.15% to 3.5% of the full similarity matrices. Higher density graphs tend to have more structural variation and thus

Edge Density ($ E / V $)	Iceberg Size, as fraction of full matrix	
	$\theta = 0.8$	$\theta = 0.9$
1	2.77%	1.47%
2	2.47%	0.63%
5	3.53%	0.15%

Table 3: Iceberg Size Relative to RoleSim Matrix

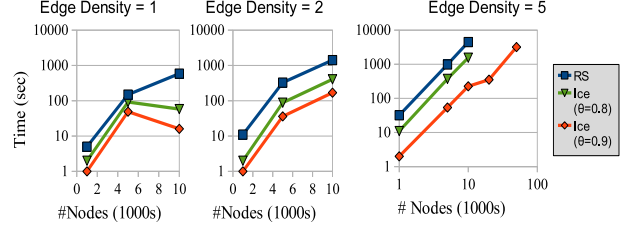


Figure 8: Execution time: Standard vs. Iceberg

fewer highly similar node pairs. In Figure 8, we see that Iceberg RoleSim is an order of magnitude faster. To check that the ranking has not changed significantly, we computed the Pearson correlation coefficient for each graph’s Iceberg RoleSim’s rankings vs. the rankings from the corresponding portion of the full similarity matrix. For $\theta = 0.8$, the average coefficient is 0.823, and for $\theta = 0.9$, it is 0.880. Both show very strong correlation, indicating Iceberg-RoleSim’s very good accuracy at ranking role-similarity pairs.

Next we fixed θ at 0.9 and varied α from 0 to 1.0 to see how sensitive is the accuracy of Iceberg RoleSim with respect to α . The results from 6 scale-free grapha are shown in Figure 9. The labels describe the number of nodes and edges of each graph. Most graphs prefer $\alpha = 0$, but some prefer a midrange value. Any value in the lower half seems acceptable.

7. RELATED WORK

The role similarity problem is a distinct special case of the more general structural or link similarity problems, which find applications in co-citation and bibliographic networks [29], recommender systems, [1] and Web search [17]. Link similarity means that two objects accrue some amount of similarity if they have similar links.

Formal definitions of role, which enable a clear idea of what is being measured, arose from the social science community [28, 37, 10]. Block partitioning can be used directly to group nodes into roles [47]. However, block modeling does not produce individual node-pair similarities. Therefore, it is not useful as a ranking method.

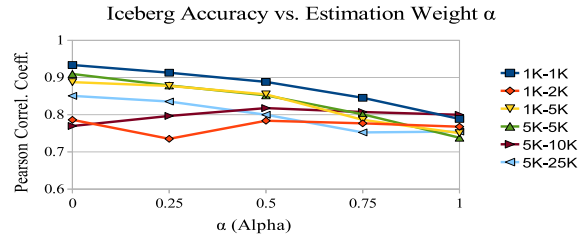


Figure 9: Iceberg Accuracy vs. α

SimRank [19] is the best known algorithm to implement a recursive definition of object similarity: two objects are similar if they relate to similar objects. SimRank has an elegant random walk interpretation: $SimRank(a, b)$ is the probability that two independent simultaneous random walkers, beginning at a and b , will eventually meet at some node. However, the more neighbors that a and b have in common, the less likely that they will both randomly choose the same neighbor. This then explains SimRank's problem with structural equivalence. Recently, Zhao [50] has pointed out that in-neighbor and out-neighbor SimRank can be used as a universal framework to describe co-citation (common in-neighbors), bibliographic coupling (common out-neighbors), or a weighted combination of the two. The number of iterations reflects the search radius for discovering similarity. As we note in Section 3.2, SimRank has an undesirable trait: its values decrease when the number of common neighbors increases. Several works have tried to address this problem. SimRank++ [1] adds a so-called *evidence* weight which partially compensates for the neighbor matching cardinality problem. In [13], they execute Monte Carlo simulations of "intelligent" random walks, where they force the overall probability of a meeting b to be Jaccard coefficient $\frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$. Recently, MatchSim [26] has also used maximal matching of neighbors to address problems with SimRank's scoring. However, our formulations have small but important differences. Because they retained SimRank's initialization, their work does not guarantee automorphic equivalence in the final results. Also, their work is intuition-based, without a theory of correctness. They provide one specific formulation, while we define a theoretical framework for *any* admissible measure or metric. Because RoleSim satisfies the triangle inequality, it is a true metric.

8. CONCLUSION

We have developed RoleSim, the first real-valued role similarity measure that confirms automorphic equivalence. We have also presented a set of axioms which can test any future measure to see if it is an admissible measure or metric. Our experimental tests demonstrate RoleSim's correctness and usefulness on real world data, opening up exciting possibilities for scientific and business applications. At the same time, we see that other well-known measures, while suitable for other tasks, are not suitable for role similarity. This axiomatic approach may prove useful for developing and validating solutions to other related tasks.

9. REFERENCES

- [1] Ioannis Antonellis, Hector Garcia-Molina, and Chi-Chao Chang. Simrank++: query rewriting through link analysis of the clickgraph (poster). In *Proc. 17th Int'l World Wide Web Conf.*, pages 1177–1178. ACM, 2008.
- [2] D. Avis. A survey of heuristics for the weighted matching problem. *Network*, 13:475–493, 1983.
- [3] V. Batagelj, P. Doreian, and A. Ferligoj. An optimizational approach to regular equivalence. *Social Networks*, 14:121–135, 1992.
- [4] Stephen P. Borgatti and Martin G. Everett. Notions of position in social network analysis. *Sociological Methodology*, 22:1–35, 1992.
- [5] Stephen P. Borgatti and Martin G. Everett. Two algorithms for computing regular equivalence. *Social Networks*, 15:361–376, 1993.
- [6] Yuanzhe Cai, Gao Cong, Xu Jia, Hongyan Liu, Jun He, Jiaheng Lu, and Xiaoyong Du. Efficient algorithm for computing link-based similarity in real world networks. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, ICDM '09*, pages 734–739, Washington, DC, USA, 2009. IEEE Computer Society.
- [7] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *PNAS*, 104(27):11150–11154, July 2007.
- [8] Dragos M. Cvetković, Michael Doob, and Horst Sachs. *Spectra of Graphs: Theory and Applications, 3rd Revised and Enlarged Edition*. Wiley, 1998.
- [9] Natalia Dragan, Michael L. Collard, and Jonathan I. Maletic. Using method stereotype distribution as a signature descriptor for software systems. In *ICSM*, pages 567–570, 2009.
- [10] Martin G. Everett. Role similarity and complexity in social networks. *Social Networks*, 7:353–359, 1985.
- [11] Martin G. Everett and Stephen P. Borgatti. Exact colorations of graphs and digraphs. *Social Networks*, 18:319–331, 1996.
- [12] M.G. Everett and S. P. Borgatti. Regular equivalence: General theory. *J. Mathematical Sociology*, 19:29–52, 1994.
- [13] Dániel Fogaras and Balázs Rácz. Scaling link-based similarity search. In *Proc. 14th Int'l World Wide Web Conf.*, pages 641–650, New York, NY, 2005. ACM.
- [14] Scott Fortin. The graph isomorphism problem. Technical Report TR 96-20, Dept. Computer Science, Univ. of Alberta, Edmonton, Alberta, Canada, July 1996.
- [15] C. Godsil and G. Royle. *Algebraic Graph Theory*. Springer-Verlag, New York, 2001.
- [16] E.M. Hafner-Burton, M. Kahler, and A.H. Montgomery. Network analysis for international relations. *International Organization*, 63(03):559–592, 2009.
- [17] Taher H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):784–796, 2003.
- [18] Petter Holme and Mikael Huss. Role-similarity based functional prediction in networked systems: application to the yeast proteome. *J. R. Soc. Interface*, 2(4):327–33, 2005.
- [19] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proc. 8th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pages 538–543, New York, NY, 2002. ACM.
- [20] Xu Jia, Yuanzhe Cai, Hongyan Liu, Jun He, and Xiaoyong Du. Calculating similarity efficiently in a small world. In *Proceedings of the 5th International Conference on Advanced Data Mining and Applications, ADMA '09*, pages 175–187, Berlin, Heidelberg, 2009. Springer-Verlag.
- [21] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963.

- [22] E. A. Leicht, Petter Holme, and M. E. J. Newman. Vertex similarity in networks. *Phys. Rev. E*, 73:026120, 2005.
- [23] Pei Li, Yuanzhe Cai, Hongyan Liu, Jun He, and Xiaoyong Du. Exploiting the block structure of link graph for efficient similarity computation. In *Proc. 13th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining*, pages 389–400, Berlin, Heidelberg, 2009. Springer-Verlag.
- [24] Zhenjiang Lin, Irwin King, and Michael R. Lyu. Pagesim: A novel link-based similarity measure for the world wide web. In *Web Intelligence*, pages 687–693. IEEE Comput. Soc., 2006.
- [25] Zhenjiang Lin, Michael R. Lyu, and Irwin King. Extending link-based algorithms for similar web pages with neighborhood structure. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 263–266. IEEE Computer Society, 2007.
- [26] Zhenjiang Lin, Michael R. Lyu, and Irwin King. Matchsim: a novel neighbor-based similarity measure with maximum neighborhood matching. In *Proc. 18th ACM Conf. Information and Knowledge Mgmt.*, pages 1613–1616. ACM, 2009.
- [27] Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, and Denis Turdakov. Accuracy estimate and optimization techniques for simrank computation. *Proc. VLDB Endowment*, 1:422–433, 2008.
- [28] F. P. Lorrain and H. C. White. Structural equivalence of individuals in networks. *J. Math. Sociology*, 1:49–80, 1971.
- [29] Wangzhong Lu, Jeannette Janssen, Evangelos Milios, and Nathalie Japkowicz. Node similarity in networked information spaces. In *CASCON '01: Proc. Conf. Centre for Adv. Studies on Collaborative research*, page 11. IBM Press, 2001.
- [30] J. J. Luczkovich, S.P. Borgatti, J.C. Johnson, and M.G. Everett. Defining and measuring trophic role similarity in food webs using regular coloration. *J. Theoretical Biology*, 220:303–321, 2003.
- [31] Ben D. MacArthur, Rubén J. Sánchez-García, and James W. Anderson. Note: Symmetry in complex networks. *Discrete Appl. Math.*, 156(18):3525–3531, 2008.
- [32] Maarten Marx and Michael Masuch. Regular equivalence and dynamic logic. *Social Networks*, 25(1):51–65, 2003.
- [33] B.D. McKay. Practical graph isomorphism. *Congressus Numerantium*, 30:45–87, 1981.
- [34] Mark Newman. Internet network.
- [35] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Univ., 1999.
- [36] Ronald Read and Derek Corneil. The graph isomorphism disease. *J. Graph Theory*, 1:339–363, 1977.
- [37] Lee Douglas Sailer. Structure equivalence: meaning and definition, computation and application. *Social Networks*, 1:73–80, 1978.
- [38] Michael Schultz and Mark Liberman. Topic detection and tracking using idf-weighted cosine coefficient. In *Proc. DARPA Broadcast News Workshop*, pages 189–192. Morgan Kaufmann, 1999.
- [39] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Amer. Soc. Information Sci.*, 24:265–269, 1973.
- [40] Malcolm K. Sparrow. A linear algorithm for computing automorphic equivalence classes: the numerical signatures approach. *Social Networks*, 15(2):151–170, 1993.
- [41] Jie Tang, Jing Zhang, Limin Yao, and Juanzi Li. Extraction and mining of an academic social network. In *Proc. 17th Int'l Conf. World Wide Web (WWW)*, pages 1193–1194, New York, 2008. ACM.
- [42] T. T. Tanimoto. An elementary mathematical theory of classification and prediction. *IBM Taxonomy Application M. and A.6*, 3, Nov. 1958.
- [43] Sudhir L. Tauro, Georgos Siganos, C. Palmer, and Michalis Faloutsos. A simple conceptual model for the internet topology. In *Proc. IEEE Global Telecommunications Conf.*, pages 1667–1671. IEEE, 2001.
- [44] Y.J. Wang and G.Y. Wong. Stochastic blockmodels for directed graphs. *J. Amer. Stat. Assoc.*, 82(397):8–19, 1987.
- [45] Stanley Wasserman and Katherine Faust. *Social network analysis: methods and applications*. Cambridge University Press, 1994.
- [46] Douglas R. White and Karl P. Reitz. Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5:193–234, 1983.
- [47] Harrison White, Scott Boorman, and Ronald Breiger. Social structure from multiple networks. i: Blockmodels of roles and positions. *Am. J. Sociology*, 81:730–780, 1976.
- [48] Wensi Xi, Edward A. Fox, Weiguo Fan, Benyu Zhang, Zheng Chen, Jun Yan, and Dong Zhuang. Simfusion: measuring similarity using unified relationship matrix. In *Proc. 28th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 130–137, 2005.
- [49] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Linkclus: efficient clustering via heterogeneous semantic links. In *Proc. 32nd Int'l Conf. Very Large Data Bases*, pages 427–438. VLDB Endowment, 2006.
- [50] Peixiang Zhao, Jiawei Han, and Yizhou Sun. P-rank: a comprehensive structural similarity measure over information networks. In *Proc. 18th ACM Conf. Information and Knowledge Mgmt.*, pages 553–562. ACM, 2009.

APPENDIX

A. PROOFS OF THEOREMS AND LEMMAS

Proof for Theorem 5 (RoleSim Convergence) Let the difference of $RoleSim(u, v)$ scores between iterations k and $(k-1)$ be $\delta^k(u, v) = RoleSim^k(u, v) - RoleSim^{k-1}(u, v)$. Also, let $D_k = \max_{(u,v)} |\delta^k(u, v)|$ be the maximal absolute difference across all u and v in iteration k . To prove convergence, we will show that D_k is monotonically decreasing, i.e., $D_{k+1} < D_k$. For any node pair (u, v) , let the maximal weighted matching between $N(u)$ and $N(v)$ computed at iteration $k+1$ be \mathcal{M}^{k+1} . Note that its weight is $w(\mathcal{M}^{k+1}) = \sum_{(x,y) \in \mathcal{M}^{k+1}} RoleSim^k(x, y)$. Without loss of generality, assume $N_u \leq N_v$, so that $\max(N_u, N_v) = N_v$ and $|\mathcal{M}| = N_u$. Given this, we observe that

$$\begin{aligned} w(\mathcal{M}^{k+1}) - (N_v \cdot D_k) &\leq \\ w(\mathcal{M}^{k+1}) - |\mathcal{M}| \cdot D_k &\leq w(\mathcal{M}^k) \leq w(\mathcal{M}^{k+1}) + |\mathcal{M}| \cdot D_k \\ &\leq w(\mathcal{M}^{k+1}) + (N_v \cdot D_k) \end{aligned}$$

Therefore, $|w(\mathcal{M}^{k+1}) - w(\mathcal{M}^k)| \leq N_v \times D_k$. Then,

$$\begin{aligned} |\delta^{k+1}(u, v)| &= |RoleSim^{k+1}(u, v) - RoleSim^k(u, v)| \\ &= |(1-\beta) \frac{w(\mathcal{M}^{k+1})}{N_v} - (1-\beta) \frac{w(\mathcal{M}^k)}{N_v}| \\ &= \frac{(1-\beta)}{N_v} |w(\mathcal{M}^{k+1}) - w(\mathcal{M}^k)| \\ &\leq \frac{(1-\beta)}{N_v} N_v \times D^k < D^k \end{aligned}$$

Therefore, $D^{k+1} = \max_{(u,v)} |\delta^{k+1}(u, v)| < D^k$, and therefore, $RoleSim^k$ will converge. \square

Proof for Lemma 4 (Triangle Inequality Invariant) For iteration k , for any nodes a, b , and c , $d^k(a, c) \leq d^k(a, b) + d^k(b, c)$, where $d^k(a, b) = 1 - RoleSim^k(a, b)$. We must prove that this inequality still holds for the next iteration: $d^{k+1}(a, c) \leq d^{k+1}(a, b) + d^{k+1}(b, c)$. To facilitate our discussion, we abbreviate $RoleSim^k(u, v)$ as $r(u, v)$, and without loss of generality, let $N_a \leq N_c$.

We utilize the following observation: *if there is a matching M between $N(a)$ and $N(c)$ which satisfies $1 - ((1-\beta) \frac{w(M)}{N_c} + \beta) \leq d^{k+1}(a, b) + d^{k+1}(b, c)$, then $d^{k+1}(a, c) \leq d^{k+1}(a, b) + d^{k+1}(b, c)$. This is because $\frac{w(M)}{N_c} \leq \frac{w(\mathcal{M})}{N_c}$, where \mathcal{M} is the maximal weighted matching between $N(a)$ and $N(c)$, and thus, $1 - ((1-\beta) \frac{w(M)}{N_c} + \beta) \geq 1 - ((1-\beta) \frac{w(\mathcal{M})}{N_c} + \beta) = d^{k+1}(a, c)$.*

In addition, we also denote the maximal weighted matching between $N(a)$ and $N(b)$ as $\mathcal{M}(a, b)$, and the maximal weighed matching between $N(b)$ and $N(c)$ as $\mathcal{M}(b, c)$. Now, we consider three cases characterizing the relationship between $N(a)$, $N(b)$, and $N(c)$.

Case 1 ($N_b \leq N_a \leq N_c$): In this case, we observe $|\mathcal{M}(a, b)| = |\mathcal{M}(b, c)| = N_b$. Given this, we consider the following matching

M between $N(a)$ and $N(c)$:

$$M = \{(x, z) | (x, y) \in \mathcal{M}(a, b) \wedge (y, z) \in \mathcal{M}(b, c)\}, |M| = N_b$$

Then, we have the following relationships:

$$\begin{aligned} &d^{k+1}(a, b) + d^{k+1}(b, c) - (1 - (1-\beta) \frac{w(M)}{N_c} - \beta) \\ &= (1-\beta) [-\frac{w(\mathcal{M}(a, b))}{N_a} - \frac{w(\mathcal{M}(b, c))}{N_c} + \frac{w(M)}{N_c}] + 1 - \beta \\ &= (1-\beta) [\frac{N_b - w(\mathcal{M}(a, b))}{N_a} - \frac{N_b}{N_a} + \frac{N_b - w(\mathcal{M}(b, c))}{N_c} - \frac{N_b}{N_c} \\ &\quad - \frac{N_b - w(M)}{N_c} + \frac{N_b}{N_c}] + 1 - \beta \\ &\geq (1-\beta) [1 - \frac{N_b}{N_a} + \frac{\sum_{(x,y) \in \mathcal{M}(a,b)} (1 - r(x, y))}{N_c} \\ &\quad + \frac{\sum_{(y,z) \in \mathcal{M}(b,c)} (1 - r(y, z))}{N_c} - \frac{\sum_{(x,z) \in M} (1 - r(x, z))}{N_c}] \\ &\geq (1-\beta) [\frac{\sum_{(x,y,z)} (d^k(x, y) + d^k(y, z) - d^k(x, z))}{N_c}] \geq 0 \end{aligned}$$

where $(x, y) \in \mathcal{M}(a, b)$, $(y, z) \in \mathcal{M}(b, c)$, $(x, z) \in M$

Case 2 ($N_a \leq N_b \leq N_c$): In this case, we observe $|\mathcal{M}(a, b)| = N_a$ and $|\mathcal{M}(b, c)| = N_b$. It follows that there is a subset $n(b)$ of $N(b)$ of size N_a that participates in both $\mathcal{M}(a, b)$ and $\mathcal{M}(b, c)$: $n(b) = \{y | (y, z) \in \mathcal{M}(b, c) \setminus \{(y, z) | \exists (x, y) \in \mathcal{M}(a, b)\}\}$. Given this, we consider the following matching M between $N(a)$ and $N(c)$:

$$M = \{(x, z) | (x, y) \in \mathcal{M}(a, b) \wedge (y, z) \in \mathcal{M}(b, c)\}, |M| = N_a.$$

Then, we have the following relationships:

$$\begin{aligned} &d^{k+1}(a, b) + d^{k+1}(b, c) - (1 - (1-\beta) \frac{w(m)}{n_c} - \beta) \\ &= (1-\beta) [-\frac{w(\mathcal{M}(a, b))}{n_b} - \frac{w(\mathcal{M}(b, c))}{n_c} + \frac{w(m)}{n_c}] + 1 - \beta \\ &= (1-\beta) [\frac{n_a - w(\mathcal{M}(a, b))}{n_b} - \frac{n_a}{n_b} + \frac{n_a - w(\mathcal{M}(b, c))}{n_c} - \frac{n_a}{n_c} \\ &\quad - \frac{n_a - w(m)}{n_c} + \frac{n_a}{n_c}] + 1 - \beta \\ &\geq (1-\beta) [1 - \frac{n_a}{n_b} + \frac{\sum_{(x,y) \in \mathcal{M}(a,b)} (1 - r(x, y))}{n_c} \\ &\quad + \frac{\sum_{(y,z) \in \mathcal{M}(b,c) \setminus \{(y,z) | \exists (x,y) \in \mathcal{M}(a,b)\}} (1 - r(y, z))}{n_c} \\ &\quad - \frac{n_b - n_a}{n_c} - \frac{\sum_{(x,z) \in m} (1 - r(x, z))}{n_c}] \\ &\geq (1-\beta) [1 - \frac{n_a}{n_b} - \frac{n_b - n_a}{n_c} \\ &\quad + \frac{\sum_{(x,y,z)} (d^k(x, y) + d^k(y, z) - d^k(x, z))}{n_c}] \end{aligned}$$

where $(x, y) \in \mathcal{M}(a, b)$, $(y, z) \in \mathcal{M}(b, c)$, $(x, z) \in m$

$$\begin{aligned} &\geq (1-\beta) [1 - \frac{n_a}{n_b} - \frac{n_b}{n_c} + \frac{n_a}{n_c}] \\ &= (1-\beta) \frac{n_b n_c - n_a n_c - n_b^2 + n_a n_b}{n_b n_c} \\ &= (1-\beta) \frac{(n_b - n_a)(n_c - n_b)}{n_b n_c} \geq 0 \end{aligned}$$

Case 3 ($N_a \leq N_c \leq N_b$): In this case, we observe $|\mathcal{M}(a, b)| = N_a$ and $|\mathcal{M}(b, c)| = N_c$. Given this, we consider the following matching M between $N(a)$ and $N(c)$:

$$M = \{(x, z) | (x, y) \in \mathcal{M}(a, b) \wedge (y, z) \in \mathcal{M}(b, c)\}$$

In addition, we define:

$$M_1 = \{(x, y) | (x, y) \in \mathcal{M}(a, b) \wedge \neg(y, z) \in \mathcal{M}(b, c)\}$$

$$M_2 = \{(y, z) | (y, z) \in \mathcal{M}(b, c) \wedge \neg(x, y) \in \mathcal{M}(a, b)\}$$

In other words, $M_1 \subset \mathcal{M}(a, b)$ and $M_2 \subset \mathcal{M}(b, c)$ do not link to each other using intermediate node $y \in N(b)$. We further denotes $m_1 = |M_1|$, $m_2 = |M_2|$, $m_3 = |M|$. Note that $m_1 = N_a - m_3$, $m_2 = N_c - m_3$, and $N_b \geq m_1 + m_2 + m_3$.

Then, we have the following relationships:

$$\begin{aligned} d^{k+1}(a, b) + d^{k+1}(b, c) - (1 - (1 - \beta) \frac{w(M)}{N_c} - \beta) &\geq \\ d^{k+1}(a, b) + d^{k+1}(b, c) - (1 - (1 - \beta) \frac{w(M)}{N_b} - \beta) &\geq \\ 1 - \beta - (1 - \beta) \left(\frac{w(\mathcal{M}(a, b))}{N_b} + \frac{w(\mathcal{M}(b, c))}{N_b} - \frac{w(M)}{N_b} \right) &= \\ (1 - \beta) \left(1 + \frac{m_3 - w(\mathcal{M}(a, b))}{N_b} - \frac{m_3}{N_b} + \frac{m_3 - w(\mathcal{M}(b, c))}{N_b} - \frac{m_3}{N_b} - \right. & \\ \left. \frac{m_3 - w(M)}{N_b} + \frac{m_3}{N_b} \right) &\geq \\ (1 - \beta) \left(1 - \frac{m_3}{N_b} + \frac{\sum_{(x, y) \in \mathcal{M}(a, b) \setminus M_1} (1 - r(x, y))}{N_b} - \frac{m_1}{N_b} + \right. & \\ \left. \frac{\sum_{(y, z) \in \mathcal{M}(b, c) \setminus M_2} (1 - r(y, z))}{N_b} - \frac{m_2}{N_b} \right. & \\ \left. - \frac{\sum_{(x, z) \in M} (1 - r(x, z))}{N_b} \right) &\geq \\ (1 - \beta) \left(1 - \frac{m_3}{N_b} - \frac{m_1}{N_b} - \frac{m_2}{N_b} + \right. & \\ \left. \frac{\sum_{(x, y, z) \in M} (d^k(x, y) + d^k(y, z) - d^k(x, z))}{N_b} \right) &\geq \\ ((x, y) \in \mathcal{M}(a, b), (y, z) \in \mathcal{M}(b, c), (x, z) \in M) & \\ (1 - \beta) \left(1 - \frac{m_1 + m_2 + m_3}{N_b} \right) &\geq 0 \end{aligned}$$

□

B. SIMRANK AND OTHER STRUCTURAL SIMILARITY MEASURES

B.1 Non-iterative Predecessors of SimRank

Bibliographical coupling [21] measures the similarity between two research publications by counting the number of works that are listed in both of their bibliographies. *Co-citation* [39] turns this around by counting the number of later works that cite both of the two original documents. As the size of a work's bibliography increases, the likelihood that it will contain a particular work increases. Therefore, a common normalization of these two measures is to divide the count by the number of distinct works cited.

We can form a *citation graph*, where each vertex is a document and a directed edge (a, b) means that document a cites document

b . Let $I(a)$ and $O(a)$ be the in-neighbor set and out-neighbor set of a , respectively. Let I_a and O_b be the in-degree and out-degree of a . Then, the normalized bibliographic coupling index is

$$S_{bc}(a, b) = \frac{|O(a) \cap O(b)|}{|O(a) \cup O(b)|}, \quad (8)$$

and the normalized co-citation index is

$$S_{cc}(a, b) = \frac{|I(a) \cap I(b)|}{|I(a) \cup I(b)|}. \quad (9)$$

These are simply the Jaccard coefficients of the out-neighbor sets and in-neighbors sets, respectively.

These two are suitable for unweighted and directed graphs. If a graph is undirected, then the two measures are the same. Suppose we have a weighted graph, though. This could be an author-collaboration graph, where edge (a, b) counts how many times author a has worked with author b . Or, it could be a bipartite document-term graph, where edge (d_a, t_b) counts the number of times that document a uses term b . Assign to each vertex a feature vector. For the homogeneous co-authorship graph, each author is a feature dimension; its feature vector is the set of edge weights to every other author. For the document-term graph, a document has a term vector, weighted according to term frequencies of the document. Then the cosine between two objects is a convenient and meaningful measure. Identical documents have cosine of 1, and documents with no features in common are orthogonal with cosine of 0.

$$S_{cos}(a, b) = \frac{A \cdot B}{\|A\| \|B\|}, \quad (10)$$

where A is the feature vector of vertex a . A small modification to the denominator, attributed to Tanimoto [42] maintains the overall behavior of the similarity function while aligning it with the Jaccard coefficient when the feature vectors are binary-valued:

$$S_{tani}(a, b) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}, \quad (11)$$

Schultz [38] adapted the well-known TF-IDF query-document similarity measure to produce a term-weighted document-document similarity measure. Here, $A(t)$ is the frequency of term t for object a , and $idf(t)$ is the inverse document frequency for term t . More generally, it is the significance or importance of term t appearing in a document.

$$S_{wcos}(a, b) = \frac{\sum_{t \in T} A(t) B(t) idf(t)}{\|A\| \|B\|} \quad (12)$$

B.2 SimRank and Simple Generalizations

Jeh and Widom [19] realized that a more general way to attack the object similarity problem was to not only look for shared neighbors, that is, neighbors that are *identical*, but to look for neighbors that are *similar*. This produces the recursive statement, "Two objects are similar if they are related to similar objects." [19] Formally, their SimRank measure is defined as follows:

$$sim_{sr}(a, b) = \frac{c}{|I(a)| |I(b)|} \sum_{x \in I(a)} \sum_{y \in I(b)} sim_{sr}(x, y) \quad (13)$$

Obviously, we can add the effects of in-neighbors and out-neighbors to produce a more comprehensive measure of the neighbor similarity between two objects. Several authors have proposed this [25, 50].

B.3 Improving the SimRank’s Computational Performance

SimRank can be described as a recursive extension of the citation index. An important difference between the non-iterative algorithms in Section B.1 and SimRank is that the earlier algorithms can be computed locally with a minimum of computational effort. With SimRank, however, to compute the similarity of even a single pair of objects, one has to consider the entire graph. This increases the computational requirements by a factor of n^2k , where k is the number of iterations. Consequently, several authors [27, 20, 6, 23] have worked to reduce both the computational and memory requirements for SimRank, for general and specific applications.

B.4 Meaningful Extensions and Alternative to SimRank

In addition to concerns about the computational efficiency of the original SimRank formula, there are some structural flaws which mar its elegance. First, SimRank scores sometimes decrease when we would intuitively expect them to increase. Suppose we have an object-pair that has all neighbors in common. Then $\text{sim}_{sr}(a, b) = c/d$, d is the degree of a or b . As d increases, this should mean stronger ties between a and b , but clearly sim_{sr} actually decreases.

B.4.1 SimRank++

Antonellis et al. [1] partially compensates for this unwanted decrease by inserting an *evidence* factor. The more neighbors in common, the higher the evidence of similarity. They define evidence as

$$ev(a, b) = \sum_{i=1}^{|N(a) \cap N(b)|} \frac{1}{2^i}, \quad (14)$$

where $N(a)$ is the undirected neighbor set of a . If a and b have only one neighbor in common, $ev = 1/2$. As the number of neighbors increases, $ev \rightarrow 1$. This yields to following similarity definition:

$$sim_{ev}(a,b) = ev(a,b) \cdot c \sum_{x=1}^{N(a)} \sum_{y=1}^{N(b)} sim_{ev}(x,y) \quad (15)$$

The very narrow range $[0.5, 1]$ of the evidence factor, however, leads to the problem that $sim_{ev}(\cdot)$ values are no longer bounded to a maximum of 1 or even to a constant. Instead, the maximum depends on the maximum value of $\|N(a)\| \cdot \|N(b)\|$ for the graph.

The authors make one more extension to support edge-weighted graphs. Their final measure is called SimRank++:

$$sim_{spp}(a, b) = ev(a, b) \cdot c \sum_{x=1}^{N(a)} \sum_{y=1}^{N(b)} w_{ab} w_{by} sim_{spp}(x, y) \quad (16)$$

B.4.2 *PSimRank*

Fogarás and Rácz [13] realize that the cause of improper weighted of neighbor-matching in SimRank is due to the paired-random walk model. Ignoring the decay constant c for the moment, SimRank values are equal to the probability that two simultaneous random walkers, starting at vertices a and b , will encounter each other eventually. Even if a and b have all $N_a = N_b$ neighbors in common, the probability that the two walkers will happen to choose the same neighbor is $1/N_a$, which decreases as the degree increases. To emend this situation, Fogarás and Rácz introduce coupled random walks. They partition the event space into three cases:

1. $P_1 = P(a \text{ and } b \text{ step to the same vertex}) = \frac{|I(a) \cap I(b)|}{|I(a) \cup I(b)|}$
2. $P_2 = P(a \text{ steps to a vertex in } I(a) \setminus I(b)) = \frac{|I(a) \setminus I(b)|}{|I(a) \cup I(b)|}$
3. $P_3 = P(b \text{ steps to a vertex in } I(b) \setminus I(a)) = \frac{|I(b) \setminus I(a)|}{|I(a) \cup I(b)|}$

Note that case 1, which we would consider the direct similarity of a and b , is described by the Jaccard Coefficient. As required, the sum of these probabilities equals 1. We can then compute a similarity measure which takes the general form

$$sim_{ps}(a, b) = \sum_{i=1}^3 P_i \cdot sim(\text{neighbors in Case } i).$$

Noting that there are $\frac{1}{|I(a) \setminus I(b)| |I(b)|}$ neighbor-pairs in Case 2 and $\frac{1}{|I(b) \setminus I(a)| |I(a)|}$ in Case 3, this produces the logical but somewhat unwieldy formula:

$$\begin{aligned} sim_{ps}(a, b) = & c \cdot [P_1 \cdot 1 \\ & + \frac{P_2}{|I(a) \setminus I(b)| |I(b)|} \sum_{\substack{x \in I(a) \setminus I(b) \\ y \in I(b)}} sim_{ps}(x, y) \\ & + \frac{P_3}{|I(b) \setminus I(a)| |I(a)|} \sum_{\substack{x' \in I(b) \setminus I(a) \\ y' \in I(a)}} sim_{ps}(x', y')]. \end{aligned} \quad (17)$$

B.4.3 MatchSim

The authors of MatchSim [26] take this emendment of random walking to its limit. They observe that when a human compares the features of two objects, a human does not select random features to see if they match. Rather, people look to see if there exists an alignment of features that produces a perfect or near-perfect matching. Therefore, their similarity measure discards the idea of random walk and replaces it with "the average similarity of the maximal matching between their neighbors." [26]:

$$sim_{ms}(a, b) = \frac{\sum_{(x,y) \in m_{ab}^*} sim_{ms}(x, y)}{\max(|I(a)|, |I(b)|)}, \quad (18)$$

where m^* represents the maximal matching. MatchSim omits the usual decay factor c , but this seems to be an idealization rather than a necessary alteration. Note that the size of the maximal matching is $\min(|I(a)|, |I(b)|)$. Without loss of generality, assume a has fewer neighbors than b . The upper bound for $\text{sim}_{ms}(a, b)$ occurs when every neighbor of a is also a neighbor of b . In this special case, $\max(\text{sim}_{ms}(a, b)) = \max(\frac{\min(|I(a)|, |I(b)|)}{|I(a) \cup I(b)|}) = \frac{|I(a) \cap I(b)|}{|I(a) \cup I(b)|}$, which is the Jaccard coefficient.

B.4.4 PageSim

All of the previous works are modifications of the original SimRank measure and principles. We now consider two measures that are markedly different than SimRank. We first consider PageSim, which not only borrows the entire PageRank computation as a starting point, but also borrows the meaning of PageRank's iterative computation to devise a related computation. The canonical interpretation of PageRank is that for each step, each page sends out an equal fraction of its own importance to each of its neighbors. Its importance for the next step is the sum of the fractional importance it received from its in-neighbors. PageSim also uses this spreading or propagating mechanism; however, rather than there being a universal importance feature which can be summed, each node begins with a distinct self-feature, which is orthogonal to every other vertex feature. The authors describe the propagation process as occurring over distinct paths, and they sum the contributions of each path to compute the total distribution. As long as we permit self-intersecting paths, this is equivalent to measuring for each vertex is the random walk distribution after k steps. PageSim follows a multi-step procedure:

1. For each vertex a , define feature vector $FV(a)$. $FV_b(a)$ is the b^{th} element of $FV(a)$.
2. Initialize all vectors: $FV_a^0(a) = \text{PageRank}(a)$. $FV_b^0(a) = 0, b \neq a$.
3. For $t = 1$ to k iterations, $FV^t = c \cdot \sum_{a \in V} \frac{FV^{t-1}(a)}{|O(a)|}$
4. Measure the similarity between pairs of feature vectors. In their original paper [24], the similarity measure is defined thus:

$$\text{sim}_{pg1}(a, b) = \sum_{i=1}^n \frac{\min(FV_i(a), FV_i(b))^2}{\max(FV_i(a), FV_i(b))} \quad (19)$$

In an expanded work [25], they modify the formula to more closely resemble the Jaccard coefficient:

$$\text{sim}_{pg2}(a, b) = \frac{\sum_{i=1}^n \min(FV_i(a), FV_i(b))}{\sum_{i=1}^n \max(FV_i(a), FV_i(b))} \quad (20)$$

B.4.5 Vertex Similarity in Networks

The last measure that we consider addresses the other major weakness of SimRank: it considers only equal-length paths of similarity. As stated earlier, a SimRank value equals the probability that a given pair of vertices will meet *if they take steps simultaneously with the other*. That is, it would not count a case

where Walker a takes 3 steps to reach c , and Walker b takes 4 steps to reach c . To address this limitation, Leicht et al. [22] formulate their measure from the following maxim: "Vertex a is similar to b if a has any neighbor c this is itself similar to b ." On one hand, this statement explicitly supports asymmetrical pairs of paths. On the other hand, it makes a questionable leap by assuming that being neighbors implies similarity.

Coming from the network science community rather than the data mining community, the authors did not give a catchy or convenient name to their measure. For convenience, we will call it VertexSim (notated sim_v or S_v). The initial version of VertexSim, written in matrix form is

$$S_v = \phi A S_v + I, \quad (21)$$

where A is the adjacency matrix and ϕ is a parameter to be determined. Solving for S_v and performing a power series expansion, we get

$$S_v = I + \phi A + \phi^2 A^2 + \dots$$

After normalizing for the expected number of paths from a to b and some simplifying approximations, they authors finally derive the following:

$$S_v = D^{-1} \left(I - \frac{c}{\lambda_1} A \right)^{-1} D^{-1}, \quad (22)$$

where λ_1 is the largest eigenvalue of A , and D is the degree matrix (d_{ii} = degree of vertex i ; all other $d_{ij} = 0$). Here we have a closed form solution, which seems convenient, but we also need to invert two matrices. An iterative computation process being simpler, the authors rewrite the equation this way:

$$D S_v D = \frac{c}{\lambda_1} A (D S_v D) + I, \quad (23)$$

which we see resembles Eq. 21. The authors claim $D S_v D$ can be initialized to any values such as $\mathbf{0}$ and will converge after 100 iterations or fewer.

B.5 Summary

We summarize the foregoing structural similarity measures in Table 4.

measure	formula
bibliographic coupling	$S_{bc}(a, b) = \frac{ O(a) \cap O(b) }{ O(a) \cup O(b) }$
co-citation	$S_{cc}(a, b) = \frac{ I(a) \cap I(b) }{ I(a) \cup I(b) }$
cosine	$S_{cos}(a, b) = \frac{A \cdot B}{\ A\ \ B\ }$
Tanimoto	$S_{tani}(a, b) = \frac{A \cdot B}{\ A\ ^2 + \ B\ ^2 - A \cdot B}$
weighted cosine	$S_{wcos}(a, b) = \frac{\sum_{t \in T} A(t)B(t)idf(t)}{\ A\ \ B\ }$
SimRank	$sim_{sr}(a, b) = \frac{c}{ I(a) I(b) } \sum_{x \in I(a)} \sum_{y \in I(b)} sim_{sr}(x, y)$
SimRank++	$sim_{spp}(a, b) = \sum_{i=1}^{ N(a) \cap N(b) } \frac{1}{2^i} \cdot c \sum_{x=1}^{N(a)} \sum_{y=1}^{N(b)} w_{ax} w_{by} sim_{spp}(x, y)$
PSimRank	$sim_{ps}(a, b) = c \cdot \left[\frac{ I(a) \cap I(b) }{ I(a) \cup I(b) } + \frac{\sum_{x \in I(a) \setminus I(b), y \in I(b)} sim_{ps}(x, y)}{ I(a) \cup I(b) I(b) } + \frac{\sum_{x' \in I(b) \setminus I(a), y' \in I(a)} sim_{ps}(x', y')}{ I(b) \cup I(a) I(a) } \right]$
MatchSim	$sim_{ms}(a, b) = \frac{\sum_{(x,y) \in m_{ab}^*} sim_{ms}(x, y)}{\max(I(a) , I(b))}$
PageSim [25]	$sim_{pg2}(a, b) = \frac{\sum_{i=1}^n \min(FV_i(a), FV_i(b))}{\sum_{i=1}^n \max(FV_i(a), FV_i(b))}$
VertexSim	$DS_v D = \frac{c}{\lambda_1} A(DS_v D) + I$

Table 4: Structural Similarity Measures